AD-A031 391 NAVAL RESEARCH LAB WASHINGTON D C   F/G 17/2
LISTENER PREFERENCE AND COMPREHENSION TESTS OF STRESS ALGORITHM--ETC(U)
SEP 76 A MCHUGH

UNCLASSIFIED   NRL-8015          NL

|OF|

AD
A031391

END
DATE
FILMED
11-76

1.0

1.1

1.25 1.4 1.6

1.8

2.0

2.2

2.5

2.8

3.2

3.6

4.0

4.5
5.0
5.6
6.3

MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS–1963–A
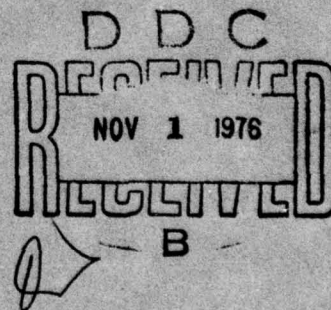
NRL Report 8015

# Listener Preference and Comprehension Tests of Stress Algorithms for a Text-to-Phonetic Speech Synthesis Program

ASTRID MCHUGH

*Information Systems Staff*
*Communication Sciences Division*

September 9, 1976

NAVAL RESEARCH LABORATORY
Washington, D.C.

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>NRL Report 8015 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE *(and Subtitle)*<br>LISTENER PREFERENCE AND COMPREHENSION TESTS OF STRESS ALGORITHMS FOR A TEXT-TO-PHONETIC SPEECH SYNTHESIS PROGRAM | | 5. TYPE OF REPORT & PERIOD COVERED<br>Interim report on a continuing NRL Problem |
| | | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>Astrid McHugh | | 8. CONTRACT OR GRANT NUMBER(s) |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Department of the Navy<br>Naval Research Laboratory<br>Washington, D.C. 20375 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>B02-15<br>Project RF21-211-401 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS | | 12. REPORT DATE<br>September 9, 1976 |
| | | 13. NUMBER OF PAGES<br>24 |
| 14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* | | 15. SECURITY CLASS. *(of this report)*<br>Unclassified |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Context-sensitive rules for speech synthesis
Intonation
Speech synthesis
Spelling-to-sound rules
Stress algorithms

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Six different stress algorithms were tested to find a relatively simple set of rules that would work well with a text-to-phonetic program of limited size. Algorithms using spelling rules only (no lexical, syntactic, or semantic information) were compared with a monotone and with hand-placed stress (English pronunciation rules). In the preference test, an algorithm that uses statistical regularities in English to assign stress and timing was judged better than a monotone, random stress, and strictly alternating stress. Hand-placed stress was judged better than this algorithm only when

(Continued)

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE
1 JAN 73
S/N 0102-014-6601

i

20. (Continued)

combined with timing rules. Comprehension tests revealed differences only with unpracticed subjects, for whom the pattern of results was similar to that of the preference tests.

| ACCESSION for | | |
|---|---|---|
| NTIS | White Section | ☑ |
| DDC | Buff Section | ☐ |
| UNANNOUNCED | | ☐ |
| JUSTIFICATION | | |

BY
DISTRIBUTION/AVAILABILITY CODES

| Dist. | AVAIL. and/or SPECIAL | |
|---|---|---|
| A | | |

ii

## CONTENTS

# LISTENER PREFERENCE AND COMPREHENSION TESTS OF STRESS ALGORITHMS FOR A TEXT-TO-PHONETIC SPEECH SYNTHESIS PROGRAM

## INTRODUCTION

The availability of commercial speech synthesizers at reasonable prices is making computer voice output a practical reality. A variety of American English text-to-phonetic programs, which convert English spelling to the phonetic input required by a synthesizer, are being developed by several groups in this country. For practical use in combination with other systems, such programs should not use an excessive portion of the computational facilities. The text-to-phonetic program developed at NRL [1] uses a limited set of context-sensitive letter-to-sound rules. It does not include a lexicon of specific words (although a few of the rules are so specific as to be essentially pronunciations of certain irregular words) and contains no syntactic information other than punctuation. The present version gives the correct pronunciation of about 90% of English words. Most of the remaining words have a single error easily corrected by the listener.

Besides acceptable pronunciation, an important feature of natural speech is variation in stress; not all sounds are given the same emphasis. It is increasingly apparent that the prosodic characteristics of speech (frequency, duration, intensity) are important to understanding. The rhythmic nature of natural speech is used by listeners to anticipate future aspects of the signal, based on past and present information [2,3]. In natural speech, stressed elements tend to be better articulated and more informative than unstressed elements. The rhythm and stress patterns of a sentence also convey information about its syntactic structure, and intonation often indicates the emotional content of the message.

The flat, machine-like sound of synthetic speech when there are no stress variations is very striking. Most people find it annoying and difficult to listen to. Adding stress variations to a text-to-speech program should make listening more pleasant and could reduce boredom and fatigue in extended listening. It is possible that almost any stress variation, even if it does not follow correct English pronunciation, would relieve the machine monotone enough to make listening easier. On the other hand, incorrect stresses might cause the listener to falsely interpret words or sentence structures. A reasonably correct stress pattern, in contrast, would probably improve comprehension and make listening more pleasant.

Some of the more sophisticated text-to-speech programs do include provisions for assigning different stress levels, particularly in polysyllabic words. Though there are regularities in English as to stress placement in words and sentences [4], there are also numerous exceptions to these regularities. Linguists disagree as to a correct set of stress rules for English. Even if a complete set of rules could be stated, it would be very complicated and unwieldy. The rules for word stress would require information about roots, prefixes,

suffixes, and part of speech; rules for sentence stress would require complex syntactic and semantic information.

A system for rule synthesis of speech which included a large body of lexical, syntactic, and semantic information might provide an excellent test of various theories of stress in English. On the other hand, many text-to-phonetic programs *do* have a little or no lexicon information and are even less likely to include syntactic rules, except those that deal with punctuation. To include such information would require extra memory and computational facilities and would limit practical uses of the program. For practical rather than theoretical uses, therefore, it is better to have a simple set of rules containing little information beyond spelling, word boundaries, and punctuation. Such unsophisticated rules would obviously lead to erroneous stress assignment in a number of cases. However, even the most complex stress rules would occasionally lead to wrong assignments.

It should be possible to define a fairly simple set of stress rules that will improve both quality and comprehension. In refining such rules, a point will be reached at which small improvements in quality and comprehension can be gained only at the expense of great complexity. This compromise between practicality and accuracy or acceptability is similar to that which motivated the development of the NRL text-to-phonetic rules. The experiments discussed in this report compare unstressed synthetic speech, simple stress rules, and two versions of hand-placed stress (stress placed by hand should represent the best that could be attained if the rules never made wrong assignments).

A discussion of some of the problems of stress in natural and synthetic speech precedes the report on the listener preference and comprehension experiments. Because of the somewhat unsatisfactory outcome of the comprehension test, a following section discusses problems of testing comprehension and gives recommendations for future tests.


## PROBLEMS OF STRESS

### Acoustic Correlates of Stress

Three acoustic characteristics are often associated with stress: fundamental frequency (or pitch), syllable duration (or length), and intensity (or loudness). In traditional linguistics, the word *stress* referred to intensity, and it was assumed that stress was perceived as an increase in loudness. Intonation (or change in pitch) was supposed to be separately perceived. Experimental evidence suggests that stress and pitch are not separately perceived and that these judgments may depend on the entire sentence contour [5,6] Bolinger [5,7] has proposed that stress in English is actually pitch accent.

Experiments by Fry [8], using words that can be heard either as a noun (SUBject) or as a verb (subJECT), showed that greater duration or greater intensity could lead to the perception of stress on a given syllable but that duration was the stronger cue. Later, similar experiments [9] confirmed this and extended the experiments to include frequency. The effect of frequency cues on noun vs verb judgments tended to be all-or-none, and frequency seemed to be the overriding cue when it was opposed to duration. Morton and Jassem [10] obtained similar results using nonsense words and found the effect of frequency to be substantially greater than that of duration or intensity. Lehiste [11]

cites studies reporting similar results for Swedish, Polish, and French. Many of the experiments on stress have dealt with isolated words rather than words in sentences. There is no particular reason to expect that these results apply to running speech. Words spoken in isolation differ in many ways from words spoken in sentences. For example, words spoken in sentences are often less clearly articulated than the same words spoken in isolation. Also, the last syllable before a pause is often lengthened. This means that the final syllable of a polysyllabic word spoken in isolation is probably longer than if the word were in the middle of a sentence.

Lieberman [12] compared stressed and unstressed syllables of words spoken in a sentence context. Stressed syllables were more often higher in frequency, longer, and of greater intensity than unstressed syllables. Umeda [13] collected vowel duration data from extended readings. In addition to differences in duration (depending on the particular vowel and the following consonant), stressed vowels tend to be longer on the average than unstressed vowels, and the last vowel in the sentence or phrase is substantially lengthened. This final lengthening may in part account for Klatt's [14] finding that the vowel is longer in a monosyllable than in a disyllable.

While stressed syllables do tend to be longer, on the average, than unstressed syllables, the relationship between stress and timing in running speech is probably more complex than this [2,15,16]. The temporal organization is influenced by the pattern of accents in the entire sentence [2]. The context of surrounding stress levels in the sentence, as well as the stress level of the syllable itself, influence the duration of a given syllable. One example of this is that [17] in natural speech stressed syllables tend to be equidistant in time (isochronous stress), so that a stressed syllable just before another stress is somewhat longer to compensate for the lack of intervening syllables, and multiple adjacent unstressed syllables will tend to be compressed. True isochrony rarely occurs, because there are limits to the amount of compensatory lengthening and shortening that actually takes place.

Although no single acoustic characteristic of the speech signal corresponds perfectly with the production or perception of stress, Lieberman [12] found no case in which the stressed syllable did not have either higher fundamental frequency or greater intensity, or both. Electrophysiological studies have led some investigators to suggest that stress corresponds most closely with physical effort [18,19]. It seems clear that the acoustic manifestations of stress are due to the interaction of a number of factors, including articulatory constraints and the context of surrounding speech.

Martin [2] proposed a speaker-listener model, in which the listener is able to use past and present information in the speech signal to anticipate the occurrence of stressed elements in real time. The perception of stress, or accent, then, depends on the interactions of temporal and other acoustic variables within the sentence. The acoustic correlates of stress are related to the rhythmic organization and internal relationships in the entire unit and need not occur "on" the actual segment that is stressed.

In summary, frequency seems the most reliable single concommitant of stress in English. The relative timing of a syllable in a sentence depends on the surrounding stresses as well as on the stress of the syllable itself. Intensity is also related to stress, but is a less reliable cue. On the segmental (or phoneme) level, vowel quality is an important cue to stress. Vowel reduction frequently occurs in unstressed syllables; for example, the /ae/

3

sound in *add* moves toward the neutral vowel schwa /ə/ in the unstressed first syllable of the word *addition*. Stress depends on the interaction of segmental cues (information about a single phoneme) and suprasegmental cues (information over more than a single phoneme).

## Introducing Stress in Synthetic Speech

In designing a text-to-phonetic translation and synthesis system, two kinds of choices need to be made in adding stress to the synthetic speech: which acoustic variables should be used to produce perceived stress and how stress levels should be assigned to different phonetic segments.

Which acoustic variations to make is considered in designing synthesizers as well as in writing programs for their use [20-22]. For example, the Votrax VS 6.0, which was used in the experiments reported here, has four choices of "inflection level." These affect both pitch and, to a lesser extent, intensity. Vowel duration can be assigned independently. The instructions suggest assigning lower inflection levels and shorter vowels to unstressed syllables. The implication seems to be that perceived stress is easily realized in the synthesized output once the stress levels are selected. The preceding discussion of the acoustic correlates of stress indicates that for entire sentences it is not that simple. Whereas none of the acoustic correlates of stress (pitch, intensity, duration) is necessarily present, in the absence of conflicting acoustic cues each can induce perceived stress on a given syllable in isolated words [8,9]. The perception of stress in whole sentences is more complex and less consistent [5,6]. On the whole, however, a change in pitch seems to be the best candidate for causing stress to be perceived on a particular syllable, and increased duration is also correlated with stress. Temporal relationships are affected by the overall rhythmic structure of the sentence as well [2]. Based on the evidence so far, a reasonably promising approach for inducing perceived stress may be to give rising pitch to the segments selected to receive stress and then to adjust timing to create a reasonably natural temporal pattern that will agree with and supplement the percept suggested by the pitch contour. (This is an oversimplification of the intonation in a sentence or phrase [5,21,23]. The falling pitch at the end of a sentence and rising pitch at the end of a question can be implemented fairly easily. Other consistencies in intonation for which consistent rules can be found should also be considered.)

The solution to the problem of where to locate stresses will depend to a large extent on the amount of information available in the text-to-phonetic program. Where lexical information is available, much can be done with word-level stress. In the absence of such information, is it possible to achieve stress consistent enough with the sounds of English to be an improvement over a machine-like monotone?

## EXPERIMENTS

### Stress Version Selected for Testing

Six different stress patterns were tested, with the goal of finding a relatively simple set of stress rules that would work well with a text-to-phonetic program of limited size.

1. MONOTONE. For comparison with the versions which did have stress variations, this *unstressed* version used the same inflection level (level 2 on the Votrax) for all phonemes, and vowel durations were not adjusted. Any attempt at adding stress that cannot be shown to yield a better sound or better comprehension than this is probably not worth the extra expense.

2. ALTERNATING STRESS and 3. RANDOM STRESS. The very noticeable monotone when there are no stress variations has led some people to suggest that almost any variation in stress, even if not always correct, would be an improvement. It is possible, however, that incorrect stresses could cause the listener to incorrectly interpret the syntax or the meaning of certain words, thereby making the speech harder to understand. Two very simple rules were used to test the possibility that any stress at all might be better than none: (a) stressing every other syllable in the sentence and (b) assigning stress at random to syllables in the sentence. (For the definition of a syllable in this context, see Appendix A.) Alternating stresses would be more predictable, but the sing-song sound might be annoying and not much better than a monotone. Random stress would avoid this problem. These rules used two inflection levels, stressed (level 3 on the Votrax) and unstressed (level 2). Increasing the contrast between stress levels by using more levels or levels further apart would be likely to emphasize the often incorrect stresses assigned by these simple rules. Timing rules were not included because further enhancing wrongly stressed syllables might be misleading to the listener and because the main reason for testing these rules was their simplicity.

4. NRL ALGORITHM. As a contrast to the very simple rules, which often assign wrong stresses and may not sound very good, a set of rules was developed that takes advantage of statistical regularities in spelling to reduce the number of errors. These rules use some of the characteristics of English to determine stress assignment, but can still be used with a text-to-phonetic program without semantic or syntactic information. Function words such as "a" and "the," and certain endings such as "-ly" and "-ing," were alway unstressed. In some cases, other syllables were stressed or unstressed, depending on regularities embodied in the particular rules used in the NRL spelling-to-phonetic program [1]. (See Appendix A.) Aside from these restrictions, stress was assigned to alternating syllables. Because there is some tendency in English for stresses to alternate, the exceptions served a "self-correcting" function, keeping this tendency aimed at the right syllables. These rules were combined with timing rules to give sentence patterns closer to natural patterns than either RANDOM or ALTERNATING stress could produce. Stressed vowels were assigned slightly longer durations than unstressed vowels, and the timing rules shortened vowels when there were several adjacent unstressed vowels and lengthened stressed vowels immediately before other stresses (tendency toward isochronous stress). In addition, the last syllable in the sentence was lengthened. Finally, a falling intonation preceded a period and a rising intonation preceded a question mark [23]. Except for the final falling or rising intonation (inflection level 1 or 4 on the Votrax), two inflection levels were used to indicate stress level (level 3 for stressed and level 2 for unstressed). This was done to maintain comparability with the other versions and to avoid further enhancing any errors in stress.

5. ENGLISH WITHOUT TIMING and 6. ENGLISH WITH TIMING. Two versions that required hand-placed stress (i.e., the rules to determine stress could not be stated easily as a computer algorithm) were tested. Acceptable stress was assigned in agreement with correct English pronunciation. (For a given printed sentence, more than one way to assign

stress may be correct.) This may be considered to be the ideal, given the best possible algorithm and all the necessary semantic and syntactic information. This extreme case and the other extreme, the monotone version, can be used to judge the success or failure of the stress rules.

To help evaluate the importance of the timing rules (which should give a somewhat more human rhythm to the sentence) two "good" versions were tested—one with timing rules and one without. ENGLISH WITH TIMING included changes in syllable duration based on stress level and on the pattern of stresses within the sentence. The choices of vowel duration were essentially similar to those described for the NRL ALGORITHM above. ENGLISH WITHOUT TIMING had exactly the same choice of inflection levels based on stress assignment, but vowel durations were not varied in accordance with stress and context. To maintain comparability with the other versions, two levels of inflection were used (Votrax level 3 for stressed and level 2 for unstressed), with levels 1 and 4 being used only for the final falling inflection ending a sentence, or rising inflection ending a question.

Two experiments were carried out to evaluate the six stress versions. The first assessed listener preferences to determine whether adding stress made the synthetic speech sound better. The second was designed to find out if those versions that sounded better were also easier to understand. While it is likely that the more natural sounding versions would also be easier to understand, it is possible that some versions that did not sound pleasant might be quite comprehensible or that a pleasing sound might not lead to better comprehension.

### Experiment I—Listener Judgments

As a first step in comparing and evaluating the stress rules, listener judgments were obtained for the six versions of synthetic speech described in the preceding section (MONOTONE, ALTERNATING, RANDOM, NRL ALGORITHM, ENGLISH WITH TIMING, and ENGLISH WITHOUT TIMING). A system that sounds better is more likely to be found acceptable by potential users. If speech sounds good, it is also likely that it will be longer before fatigue and boredom become a problem in listening.

*Method*—Twelve sentences, selected primarily from newspapers and magazines, were translated to phonetic symbols using the NRL text-to-phonetic program. The sentences are listed in Appendix B. Obvious mispronunciations were corrected so that any deficiencies in the translation program would not influence the judgments of the six versions. Changes were made as necessary in inflection level (this varies pitch and intensity), and in vowel duration for the appropriate phonemes in the phoneme string to create six versions of each sentence (MONOTONE, ALTERNATING, RANDOM, NRL ALGORITHM, ENGLISH WITH TIMING, and ENGLISH WITHOUT TIMING). All six versions consisted of the same sequence of phonemes, and only the choice of inflection level and vowel duration differed among versions. ENGLISH WITH TIMING and ENGLISH WITHOUT TIMING were exactly alike except for certain vowel durations. The resulting 72 sentences were synthesized, using the Votrax VS 6.0, and were recorded on tape.

Each of 17 volunteers listened to the different versions in random order and judged the "goodness" of each. A seven-point scale was used, and the anchor points were labeled

simply "Good" and "Bad." Since the listeners heard each of the 12 sentences six times, they were given typewritten copies of the sentences to minimize any tendency to judge later repetitions (which otherwise might be better understood) as sounding better.

*Results*—The listeners varied in their use of the seven-point scale. Some used mainly the lower end of the scale, some used the entire scale, and some avoided the extremes and used only the middle of the scale. A ranked order of the six versions was made for each listener's judgments, in order to obtain scores that could be compared across listeners. That is, the ratings for all twelve sentences heard by the listener for a given version were used to compute a mean goodness score for that version for that listener. The six means (one for each version) for each individual listener were then ranked from highest to lowest and assigned ranks from one to six. The ranks for the different versions were remarkably consistent across listeners and fell in three groups: (a) ENGLISH WITH TIMING was significantly better than all of the others; (b) ENGLISH WITHOUT TIMING and the NRL ALGORITHM did not differ significantly from one another, but were judged significantly better than the remaining three versions; and (c) RANDOM STRESS, ALTERNATING STRESS, and the MONOTONE did not differ significantly from one another. (The Wilcoxon matched-pairs signed ranks test [24] was used for the statistical comparisons.*) The mean rank for each version was as follows.

| *Version* | *Mean Rank* |
|---|---|
| ENGLISH WITH TIMING | 1.35 |
| ENGLISH WITHOUT TIMING | 2.88 |
| NRL ALGORITHM | 2.94 |
| RANDOM | 4.53 |
| ALTERNATING | 4.65 |
| MONOTONE | 4.65 |

*Conclusions*—The two versions with very simple rules, ALTERNATING and RANDOM, were no better than MONOTONE. Both placed stress on a number of syllables or words that would not ordinarily be stressed in English. Several listeners reported that this was "distracting" or "sounded wrong." It would seem that adding some stress variation does not make the speech sound better than a monotone if stress occurs too often in the wrong places. It is possible that the stress variations improved the sound of the speech but that this effect was canceled by the negative effect of incorrect stresses.

The algorithmic version included both heuristics for stressing wrong syllables less often and some simple timing rules. This apparently made this version sound better than all but the two versions with hand-placed stress. The version with hand-placed, correct stresses indicated by pitch and intensity but without timing (ENGLISH WITHOUT TIMING) was no better than the algorithm that used timing but assigned some incorrect stresses. This fact, together with the fact that the version with hand-placed stress with timing was better than either of these, suggests that the timing of the syllables is important to the sound of the

___
*Critical values for $p < 0.05$ were used for all tests of statistical significance, and can be assumed in all cases where statistical significance is mentioned.

synthetic speech. Timing rules should probably be included in a stress algorithm for synthesis using a text-to-phonetic program.

It is encouraging that a relatively simple set of stress rules, with no syntactic or semantic information, produced synthetic speech that sounded better than a monotone. The version with hand-placed inflection and timing sounded even better. There is no agreed-upon set of rules that correctly specifies stress placement and syllable timing in English sentences, so that even if the program included the proper syntactic information (and perhaps also semantic information) it is not certain that a completely correct stress algorithm could be written. Simple stress rules can improve the sound of the speech, but better rules could improve it still more.

### Experiment II—Comprehension Test

The results of the listening test suggest that relatively simple stress rules can be added to a text-to-speech synthesis program to make it sound better. This should make the speech more pleasant to listen to and may reduce fatigue and boredom. On the other hand, listening judgments can not be used to indicate whether the speech is more intelligible. Naturalness and intelligibility do not always go together, and in some cases, it *seems that one is attained at the expense of the other* [25].

The next experiment was designed to compare comprehension for the six versions. Intelligibility is often used to refer to the identification of words or contrasting phonemes when single words are heard in isolation. Comprehension will be used here to refer to the degree to which longer passages—sentences or paragraphs—can be understood. It is more difficult to obtain a reliable measure of comprehension, but it may be a more meaningful measure of the performance of a system in actual use. The Diagnostic Reading Scales,* a reading comprehension test designed for grade school children, was selected because it was an already existing test with standardized questions and answers. It was designed to test comprehension, although in this case the norms regarding the grade level of achievement were not relevant. In the light of the final outcome, it appears that a straightforward comprehension test may be too sensitive to individual differences in ability to be useful for comparing small differences in quality.

*Method*—The subjects, 145 undergraduates at the University of Maryland who volunteered to participate for extra credit in psychology courses, heard eight tape-recorded paragraphs (two practice paragraphs and six test paragraphs) from the Diagnostic Reading Scales. After hearing each paragraph, the subjects wrote answers to a standard set of questions. Appendix C gives one of the paragraphs used in the test. The subjects were tested in groups varying in size from one to ten. All eight paragraphs heard by any one group were made using the same stress version. The six versions of each paragraph were synthesized and recorded in the same way as the materials in Experiment I. In addition, a seventh version was recorded with a human male reading the same eight paragraphs. After the last test, the subjects were asked to write any comments they had about the sound of the speech on the back of the booklet.

---

*Adapted from the Diagnostic Reading Scales devised by George D. Spache for use by the Naval Research Laboratory, Washington, D.C., with the permission of the publisher, CTB/McGraw-Hill, Del Monte Research Park, Monterey, Calif. 93490. Copyright 1963, 1972 by McGraw-Hill, Inc.

*Results*—The total number of correct answers to the questions based on the six test paragraphs was scored for each subject. The mean scores and variances for the seven apes are listed below. Human speech was not noticeably more comprehensible than the synthetic versions. The differences between means were relatively small, and the variances of subjects within groups were large. That is, individual differences in performance on the comprehension test were so large as to obscure any smaller differences in tape versions.

| Tape Version | Mean Comprehension Score | Variance |
|---|---|---|
| MONOTONE | 38.8 | 38.2 |
| ALTERNATE | 38.8 | 37.5 |
| RANDOM | 39.3 | 52.8 |
| ALGORITHM | 39.6 | 32.1 |
| ENGLISH WITHOUT TIMING | 41.6 | 26.1 |
| ENGLISH WITH TIMING | 42.0 | 16.8 |
| HUMAN SPEECH | 40.4 | 32.3 |

The maximum possible score was 52. An analysis of variance showed no statistically significant differences in comprehension scores for the seven versions, $F(6,138) = 1.08$.

The practice paragraphs were originally intended to make it easier for the listeners to understand the synthetic speech. When it appeared that the test paragraphs were perhaps too understandable, the practice paragraphs were scored as well.

An analysis of the scores for the two practice paragraphs yielded the results shown in Fig. 1. An analysis of variance showed overall statistical significance, $F(6,144) = 8.84$. The Newman Keuls test [26] was used to test differences between pairs of means; the results are shown in Table 1. Human speech was understood better than the MONOTONE, NRL ALGORITHM, ALTERNATE, and RANDOM. ENGLISH WITHOUT TIMING and ENGLISH WITH TIMING were both better than the MONOTONE, ALTERNATE, and RANDOM. The NRL ALGORITHM was better than ALTERNATE and RANDOM.

Comments made by listeners are summarized in Table 2. These comments were made in response to a request for general comments rather than as answers to particular questions. They are more likely to reflect what a particular listener thought was important or striking about the speech than what the same person might have said if asked a specific question. The comments may be most useful for constructing a questionnaire for future tests. Comparisons between the different tape versions should be made with extreme caution. It can be seen, however, that comments about the synthesized versions are very different from those made about the human speech. It is also interesting to note the number of "too fast" responses. A common complaint of people who are unfamiliar with a language is that native speakers of the language talk too fast. When greater effort is necessary for understanding speech, there seems to be a tendency to think the speech is
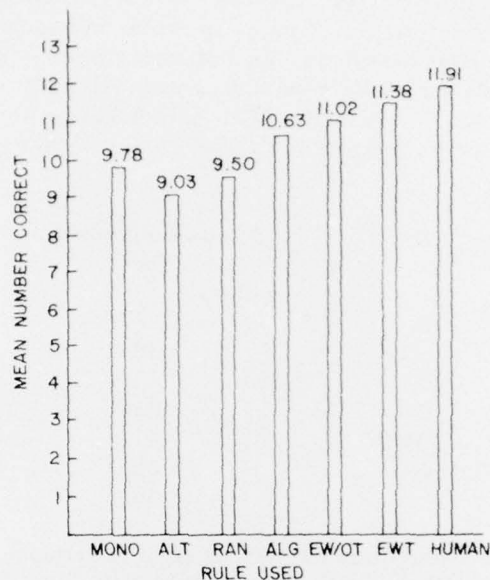
Fig. 1 — Mean scores for practice tests. The
highest possible score is 14.

going too fast. In contrast to the comments given, the NRL ALGORITHM and ENGLISH
WITH TIMING were actually slightly faster than the other versions due to the shortening
of many unstressed vowels. Among the synthetic versions, there seem to be more differ-
ences among tapes in comments on the quality of the speech (1-11) than on its difficulty
(12-18). This reflects the results of Experiment I (significant differences in quality) and
Experiment II (nonsignificant differences in comprehension).

*Discussion and Conclusions*—Human listeners are highly adaptable processors of speech
sounds and do reasonably well at understanding even badly distorted speech. Also, lan-
guage is highly redundant, so that it is often possible to fill in parts that have been missed,
based on the surrounding context. It would seem that the brief practice was enough to
enable most people to follow even the poorly rated versions of synthetic speech. The
fact that there were insignificant differences in comprehension between synthetic and
human speech when the listeners were practiced is encouraging for future users of synthe-
tic speech. Ainsworth [27] also reports high intelligibility scores for synthetic speech
with practiced listeners. This agrees with the experience of people who use synthesizers
regularly and feel that with practice synthetic speech becomes quite understandable. This
result in the present case is based on null findings and should be viewed with caution. It
may depend on the particular test circumstances. The materials were relatively easy; the
average reading level for the paragraphs was early fourth grade. There were relatively few
distractions in a reasonably quiet room, and the listeners had no other task than to listen
to the speech knowing that they were going to be questioned on the content of the para-
graphs. It is possible that if conditions were changed (for example, more distractions or a
more exacting listening task) there would be greater differences in performance.

Table 1 — Differences Between Pairs of Means on the Practice Portion of the Comprehension Test

| Tape Version | MONOTONE (9.78)† | ALTERNATE (9.03) | RANDOM (9.50) | ALGORITHM (10.63) | ENGLISH WITHOUT TIMING (11.02) | ENGLISH WITH TIMING (11.38) | Human Speech (11.91) |
|---|---|---|---|---|---|---|---|
| MONOTONE | — | -0.75 | -0.28 | 0.85 | 1.24* | 1.60* | 2.13* |
| ALTERNATE | | — | 0.47 | 1.60* | 1.99* | 2.35* | 2.88* |
| RANDOM | | | — | 1.13* | 1.52* | 1.88* | 2.41* |
| ALGORITHM | | | | — | 0.39 | 0.75 | 1.28* |
| ENGLISH WITHOUT TIMING | | | | | — | 0.36 | 0.89 |
| ENGLISH WITH TIMING | | | | | | — | 0.53 |

*These differences were statistically significant. Critical values for $p < 0.05$ were used for all comparisons—Newman Keuls test for paired comparisons [26].

†The maximum possible score was 14. Mean scores are shown in parentheses.

Table 2 — Summary of Comments*

| Comments | MONOTONE | ALTERNATING | RANDOM | NRL ALGORITHM | ENGLISH WITHOUT TIMING | ENGLISH WITH TIMING | Human Speech |
|---|---|---|---|---|---|---|---|
| 1. Unpleasant/dull, boring/annoying after a while/ didn't like it | 0.44 | 0.31 | 0.19 | 0.22 | 0.20 | 0.08 | — |
| 2. Impersonal/monotone/no expression/robot | 0.44 | 0.13 | 0.31 | 0.13 | 0.05 | 0.08 | 0.06 |
| 3. Choppy/not smooth | — | 0.13 | — | 0.09 | 0.25 | 0.05 | 0.06 |
| 4. Wrong stresses/fluctuating tone | — | 0.13 | 0.25 | 0.16 | 0.10 | 0.08 | — |
| 5. Syllables odd (e.g., some syllables too well pronounced or some syllables slurred). | 0.22 | — | 0.19 | 0.13 | 0.10 | — | — |
| 6. Sentence or words run together/no pauses/or wrong pauses | 0.18 | 0.31 | 0.31 | 0.09 | 0.15 | 0.12 | — |
| 7. Pronunciations or letter sounds were odd or wrong | 0.06 | 0.06 | — | 0.09 | 0.15 | 0.16 | — |
| 8. Foreign accent | 0.11 | 0.06 | — | 0.19 | 0.10 | 0.24 | — |
| 9. Different from normal familiar speech | — | 0.13 | 0.13 | 0.06 | — | 0.04 | — |
| 10. Pauses between sentences distracting | — | — | — | — | — | — | 0.06 |
| 11. Enunciation too clear to be natural | — | — | — | — | — | — | 0.06 |
| 12. Hard to understand/blurry/unclear/garbled | 0.06 | 0.31 | 0.38 | 0.22 | 0.25 | 0.12 | — |
| 13. Hard at first, but easier with practice | 0.33 | 0.19 | 0.38 | 0.22 | 0.30 | 0.32 | — |
| 14. Some words (parts) hard or impossible to understand | 0.28 | 0.25 | 0.25 | 0.19 | 0.30 | 0.32 | — |
| 15. More confusing at end (stories harder) or with long or complex sentences | 0.11 | 0.13 | 0.06 | 0.09 | 0.05 | 0.16 | — |
| 16. Requires concentration, constant attention/ concentrate on words at expense of comprehension or memory | 0.22 | 0.25 | 0.19 | 0.09 | 0.25 | 0.24 | — |
| 17. Too fast | 0.11 | 0.06 | 0.13 | 0.03 | — | — | — |
| 18. Mostly clear, understandable, or good | 0.22 | 0.13 | 0.06 | 0.13 | 0.35 | 0.32 | 0.75 |
| 19. Problems unrelated to the speech (e.g., tired, noisy neighbor) | — | — | — | — | — | — | 0.19 |

*The numbers in the table refer to the proportion of people hearing a particular tape who made a comment similar to the one listed. Comments such as "very interesting," "good experiment," or "like a science-fiction movie" were not included.

12

The fact that there were differences on the practice tests indicates that inexperienced listeners, at least, have less trouble understanding synthetic speech the more it resembles natural speech patterns. The two versions with many incorrect stresses were harder to understand than the algorithm or the two versions with hand-placed stress, each of which had more correct stresses. Whereas the algorithm was not significantly better than the monotone, both versions with hand-placed stress were. This suggests that adding stress to synthetic speech can improve comprehension if the stress pattern is not too different from natural speech. On the other hand, the two versions with hand-placed stress did not differ significantly from the algorithm, which suggests that the algorithmic version was not much harder to understand than the versions with optimal stress placement. If consistencies in the spelling-to-phonetic rules can be found which would further reduce the number of wrong stresses, an algorithm for adding stress to synthetic speech using simple rules similar to those reported here may be promising when lexical and syntactic information is not available.

The comprehension test used was probably too sensitive to individual differences in ability or motivation (scores ranged from 21.5 to 49.5) to reflect the relatively smaller differences in comprehension. A different measure of comprehension is needed if small differences are to be detected.

## RECOMMENDATIONS FOR FURTHER COMPREHENSION TESTING

Most commonly used intelligibility tests involve the identification of words or sentences produced by the system to be tested. Intelligibility is usually measured as the percent of words correctly identified. (In the case of sentences, either all of the words or only key words may be counted.) Intelligibility in the presence of a masking noise may also be tested. In the experiment reported here, paragraphs rather than sentences were used, because one of the reasons for adding correct stress is to make it easier to listen to extended passages. Also, a longer passage would be more likely to reveal both good and bad effects of a particular stress rule.

The following comments summarize the advantages and disadvantages of the Diagnostic Reading Scales as a test of comprehension:

• The use of longer passages more closely reflects uses to which the system might actually be put.

• On the face of it, the test seems to measure the feature of the speech that is of interest, namely comprehension.

• Variability among listeners was so great as to mask possible differences in speech versions.

• The test was probably not sensitive enough to differences in the difficulty of understanding the paragraphs and too sensitive to other factors such as general intelligence or memory.

Phonetically balanced word lists and rhyme tests such as the Diagnostic Rhyme Test (DRT) by W. D. Voiers [28] are often used for comparing relatively small differences in

the intelligibility of speech synthesis and voice transmission systems. These are primarily tests of how well individual phonemes can be distinguished. A major advantage of such tests is that scores are consistent and highly reliable, so that results for different systems can readily be compared. For the development of synthesizers and voice transmission systems, where the quality of the phonetic information transmitted is important, such tests may be good predictors of how well longer passages can be understood. The intelligibility of the Votrax VS 6.0, which was used for the tests described in this report, is 74% as measured by the DRT. In sentence contexts, actual intelligibility is likely to be higher.

Words spoken in isolation are different from words in continuous speech. When short exerpts from continuous speech are presented to listeners, only about 50% of complete words are understood [29], and when words spoken in isolation are concatenated, the resulting speech is not very intelligible [25]. Superimposing the proper intonation contour on telephone numbers obtained by splicing together numbers spoken in isolation, however, is sufficient for high quality synthesis [30] and a similar process may work for sentences [21]. It is probable that different cues are important in understanding isolated words, as distinguished from continuous speech. Complete phonetic information is more important for isolated words, whereas interpretation of continuous speech relies heavily on prosodic information.

Since it is likely that a text-to-phonetic program will be used for extended passages, it is important to assess comprehension for synthesized texts longer than single words. It is not necessarily true that the accuracy to which phonemes or single words can be identified will be a good predictor of comprehension for longer passages. Phoneme and word tests will in any case be inappropriate for comparing stress algorithms for sentence-length materials.

In spite of the higher reliability of word tests, comparison of stress rules requires materials of at least sentence length. Because individual differences accounted for so much variance in the comprehension experiment described in this report, it is important to reduce this variability by testing each person on all of the speech versions of interest. Since text materials (sentences or paragraphs) also vary in difficulty, each text should be tested in all speech versions. This means that the experimental materials must be carefully counterbalanced across people. Such counterbalancing is easier if the test consists of a large number of short passages rather than a few longer passages. It is recommended that sentences rather than paragraphs be tested, even though paragraphs may be more representative of uses for the system.

In the comprehension tests of the stress algorithms, it is possible that when the speech was more difficult to understand, people simply concentrated harder. In fact, several people mentioned the strain of having to concentrate in order to understand. It is apparent from the tests so far that the synthesized speech can be understood. It is also probable on the basis of performance on the practice tests that some versions were more difficult to understand than others.

To measure differences in comprehensibility, one could make the demands on the listener so great that he can no longer just concentrate harder to understand difficult versions. Alternatively, one could use a measure that reflects the processing effort devoted to understanding the speech.

A commonly used and relatively simple technique is to degrade the signal with noise to make understanding more difficult. Either the percent intelligibility at a given signal-to-noise ratio or the signal-to-noise ratio required for a given level of intelligibility—for example 50%—can be used to compare different systems. A test of this sort could be used to compare the intonation algorithms in future comprehension tests. The decision to use a different test in the experiments reported here was based in part on the nature of the variations made in the speech signal. It is possible that adding noise results in a loss of phonetic information, and the phonetic contents of all the passages tested were identical. Superficially, a direct test of comprehension seemed to be more likely to reflect the kinds of differences that were of interest.

Tests using a second task to be done while listening to and trying to understand the speech can also be used to increase the demands on the listener. If performance is measured on the concurrent task as well as on the speech task, poorer performance on either task may be used to indicate that the speech was more difficult to understand. The following criteria should be considered in selecting a concurrent task:

- It should require mental processing (i.e., not finger tapping).

- It should require fairly continuous attention.

- Performance should not improve greatly with practice.

- It should not involve incompatible responses (such as talking while also listening).

If the speech test involves listening to 30 to 60 different sentences, a possible concurrent task is counting backwards. At a specified time (e.g., 5s) before each sentence is given, a three-digit number is given. The person is required to count backwards by threes from that number. The answers are to be written down while he is listening to the sentence, and this is to continue until a signal is given to write down the sentence (perhaps 15s after the end of the sentence). Both the number of counting responses and the accuracy of sentence recall can easily be assessed. Other possible tasks might be monitoring a visual display for a specified event, reading dials, or even solving pencil mazes.

It is possible to invent a concurrent task which requires little in the way of specialized equipment and places greater demands on the subject. However, it is common experience among psychologists that experimental subjects are incredibly adept at devising strategies to evade the demands of the task by doing something different from what was originally intended. Reitman [31] has suggested that participants in an experiment tend to view the experimental situation as a problem to be solved. The experimenter finds himself testing strategies for performing the assigned task rather than the ability for which he designed his task.

Reaction-time tasks are most likely to be successful in measuring small differences in difficulty of understanding speech. Reaction time to a phoneme-monitoring task has been used to measure small differences in processing during ongoing speech in a variety of situations. In the usual reaction time experiement, the person listens to a sentence and is asked to monitor that sentence for some target, usually a previously specified

phoneme which occurs only once in the sentence. The listener is required to press a button as soon as he hears the target. The assumption is that the more attention or processing effort needed for understanding at the time the target occurs, the slower the reaction time to that target will be. This technique has been used successfully to show that reaction time is faster when a common word is used than when a more unusual synonym is used in the same sentence [32], that reaction time to stressed syllables is faster than to unstressed syllables [33], and that reaction times to temporally displaced targets are slower than to on-time targets when tape splicing is used to change the timing of speech [34].

A good description of the techniques for running reaction time experiments can be found in Foss [32] and Savin and Bever [34]. Since properly conducted reaction time tests are sensitive to very small differences in processing difficulty, it is strongly recommended that NRL acquire the facilities to conduct reaction time tests. If there is continued interest in speech transmission systems, there will be many occasions when it will be important to detect small differences in intelligibility, but word tests may be inappriate or inadequate.

Another test that may be useful in measuring the amount of "effort" required to understand speech is the Free Conversation Test [36]. Although it is inappropriate for testing text-to-phonetic programs (due to the necessity for typing each message into the computer), it has been used successfully by the British for comparing speech transmission systems. A reaction time test, however, is preferable in that it is sensitive to differences in *the difficulty of processing the speech at the time the speech is being heard rather than to* memory effects or judgments of difficulty after the speech has been understood.

## REFERENCES

1. H. S. Elovitz et al., "Automatic Translation of English Text to Phonetics by Means of Letter-to-Sound Rules," NRL Report 7948, Jan. 21, 1976.
2. J. G. Martin, "Rhythmic (hierarchical) versus Serial Structure in Speech and Other Behavior," *Psychol. Rev.* **79**, 487-509 (1972).
3. J. G. Martin, "Rhythmic Expectancy in Continuous Speech Perception," in *Dynamic Aspects of Speech Perception*, (A. Cohen and S. Nooteboom, eds.) Proceedings of a Symposium, Eindhoven, Holland, Aug. 1975; Springer-Verlag, Berlin, 1975.
4. N. Chomsky and M. Halle, *The Sound Pattern of English*, Harper and Row, New York, 1968.
5. D. L. Bolinger, "A Theory of Pitch Accent in English," *Word* **14**, 109-149 (1958).
6. P. Lieberman, "On the Acoustic Basis of the Perception of Stress by Linguists," *Word* **21**, 40-54 (1965).
7. D. L. Bolinger, "Pitch Accent and Sentence Rhythm," in I. Abe and T. Kanekiyo, eds., *Forms of English*, Tokyo, 1965.
8. D. B. Fry, "Duration and Intensity as Physical Correlates of Linguistic Stress," *J. Acoust. Soc. Amer.* **27**, 765-768 (1955).
9. D. B. Fry, "Experiments in the Perception of Stress," *Lang. Speech* **1**, 126-152 (1958).
10. J. Morton and W. Jassem, "Acoustic Correlates of Stress," *Lang. Speech* **8**, 159-181 (1965).
11. I. Lehiste, *Suprasegmentals*, MIT Press, Cambridge, Mass., 1970.

12. P. Lieberman, "Some Acoustic Correlates of Word Stress In American English," *J. Acoust. Soc. Amer.* **32**, 451-453 (1960).

13. N. Umeda, "Vowel Duration in American English," *J. Acoust. Soc. Amer.* **58**, 434-445 (1975).

14. D. H. Klatt, "Interaction Between Two Factors that Influence Vowel Duration," *J. Acoust. Soc. Amer.* **54**, 1102-1104 (1973).

15. A. W. F. Huggins, "On the Perception of Temporal Phenomena in Speech," *J. Acoust. Soc. Amer.* **51**, 1279-1290 (1972).

16. D. K. Oller, "The Effect of Position in Utterance on Speech Segment Duration in English," *J. Acoust. Soc. Amer.* **54**, 1235-1247 (1973).

17. D. Abercrombie, *Studies in Phonetics and Linguistics*, Oxford University Press, London, 1965.

18. I. Fonagy, "Electrophysical and Acoustic Correlates of Stress and Stress Perception," *J. Speech Hearing Res.* **9**, 231-244 (1966).

19. P. Ladefoged, *Three Areas of Experimental Phonetics*, Oxford University Press, London, 1967.

20. I. G. Mattingly, "Synthesis by Rule of Prosodic Features," *Lang. Speech* **9**, 1-13 (1966).

21. J. P. Olive, "Fundamental Frequency Rules for the Synthesis of Simple Declarative English Sentences," *J. Acoust. Soc. Amer.* **57**, 476-482 (1975).

22. L. R. Rabiner, H. Levitt, and A. E. Rosenberg, "Investigation of Stress Patterns for Speech Synthesis by Rule," *J. Acoust. Soc. Amer.* **45**, 92-101 (1969).

23. P. Lieberman, *Intonation, Perception, and Language*, MIT Press, Cambridge, Mass., 1967.

24. S. Siegel, *Non-Parametric Statistics*, McGraw-Hill, New York, 1956.

25. A. N. Stowe and D. B. Hampton, "Speech Synthesis with Prerecorded Syllables and Words," *J. Acoust. Soc. Amer.* **33**, 810-811 (L) (1961).

26. B. J. Winer, *Statistical Principles in Experimental Design*, 2d ed., McGraw-Hill, New York, 1971.

27. W. A. Ainsworth, "Performance of a Speech Synthesis System," *Int. J. Man Mach. Stud.* **6**, 493-511 (1974).

28. W. D. Voiers, A. D. Sharpley, and C. J. Hehmsoth, "Research on Diagnostic Evaluation of Speech Intelligibility," AFCRL-72-0694, Jan. 24, 1973.

29. I. Pollack and J. M. Pickett, "The Intelligibility of Excerpts from Conversation," *Lang. Speech* **6**, 165-171 (1963).

30. J. P. Olive and L. H. Nakatani, "Rule-Synthesis of Speech by Word Concatenation: A First Step," *J. Acoust. Soc. Amer.* **55**, 660-666 (1974).

31. W. Reitman, "What Does It Take to Remember?," in D. A. Norman, ed., *Models of Human Memory*, Academic Press, New York, 1970.

32. D. J. Foss, "Decision Processes During Sentence Comprehension: Effects of Lexical Item Difficulty and Position Upon Decision Times," *J. Verb. Learning Verb. Behav.* **8**, 457-462 (1969).

33. J. L. Shields, A. McHugh, and J. G. Martin, "Reaction Time to Phoneme Targets as a Function of Rhythmic Cues in Continuous Speech," *J. Exp. Psychol.* **102**, 250-255 (1974).

34. R. H. Meltzer et al., "Reaction Time to Temporally Displaced Targets in Continuous Speech," *J. Exp. Psych: Human Perception and Performance* **2**, 277-290 (1976).

35. H. B. Savin and T. G. Bever, "The Nonperceptual Reality of the Phoneme," *J. Verb. Learning Verb. Behav.* **9**, 295-302 (1970).

36. L. S. Butler and L. Kiddlo, "The Rating of Delta Sigma Modulating Systems with Constant Errors, Burst Errors, and Tandem Links in a Free Conversation Test Using the Reference Speech Link," Signal Research and Development Establishment (SRDE), Christchurch, Hants., U.K., Report No. 69014, Feb. 1969.

## APPENDIX A

## A BRIEF SUMMARY OF THE NRL STRESS ALGORITHM

Stress, for the purposes of these rules, is realized primarily as a rise in frequency and a slight concommitant increase in intensity. (Inflection levels on Votrax). Syllable duration is determined in a separate operation. The following rules apply to the phonetic output and not to the original English spelling.

1. Beginning at the final punctuation, use falling intonation on the last few phonemes before a period, semicolon, or colon, and rising intonation before a question mark.

2. A "syllable" is arbitrarily defined as a single vowel (after translation to phonetic symbols, not in English spelling), and any succeeding consonants that do not belong to the next syllable. The consonant immediately preceding a vowel belongs to that syllable, and all consonants at the beginning of the word belong to the first syllable.

3. Certain syllables are never stressed. These include (a) function words such as *a*, *the*, *by*, *for in*, etc.; (b) certain endings such as *-ing*, *-ed*, *-er*, final *-y*, etc.; and (c) any syllable in which the vowel is translated as the phoneme schwa /ə/.

4. If certain rules in the translation program apply consistently in cases where the syllable is nearly always stressed or nearly always unstressed, the stress level in these cases could be assigned as part of the translation rules. The rule must be very consistent to allow this, because this tends to exaggerate any pronunciation errors. (This step was not included in the experimental versions of the algorithm.)

5. The syllable before the final punctuation is stressed if it is not an exception in 3 or 4 above.

6. Going right to left, a syllable is stressed if it is immediately to the left of an unstressed syllable and if it is not an exception; otherwise, it is unstressed.

## APPENDIX B

### SENTENCES USED IN THE PREFERENCE TESTS

1. Sugar and egg prices here continued to decline.

2. Good weather means more children playing outside.

3. The witness himself is a major source of unreliability.

4. It is hardly necessary to point out that dramatic changes have taken place.

5. Have you ever wanted to have the last word?

6. The spokesmen said Washington could not legally appoint a substitute.

7. Such elemental gliders can soar as well as glide.

8. Through the middle of the valley flowed a winding stream.

9. Because of its hardness, this steel is used principally for making razors.

10. It is a sensible engineering survey.

11. I spotted a lady downtown with lots of earrings.

12. The instrument arrived at the museum here a few days ago.

## APPENDIX C

## A SAMPLE PARAGRAPH FROM THE DIAGNOSTIC READING SCALE*

Yesterday Bob took a trip to a city market that was somewhat like a store but a great deal bigger. It didn't have any bread or canned goods like the grocery stores. But there were a great many big boxes of vegetables and fruits.

Bob was hungry and wanted just one plum or cherry to taste. He wondered if one of the men would sell him just one plum. Everyone was buying the fruit and vegetables by the whole crate. When Bob asked the man to sell him one plum, he laughed and gave Bob an extra large plum wrapped in paper but wouldn't take any money.

As he walked along eating the plum, Bob watched the men unloading the trucks and big trailers. They would chop open the top of the crate so that anyone could see the fruit. If a buyer liked the fruit, and was willing to pay the price, he might buy the entire truckload.

Questions (Correct answers in parentheses):

1. What is a city market like? (store, but bigger or big store or bigger than a store)

2. What does the grocery store have that the city market doesn't have? (Bread or canned goods)

3. What did he ask the man for? (plum)

4. What did the man do? (gave the plum to him or laughed)

5. How much did Bob pay for the plum? (nothing)

6. What were the men doing to the trucks and trailers? (unloading)

7. Why did they open the crates? (so anyone could see the fruit)

8. If a man liked the fruit, what might he do? (ask price or buy it)

---